Distributed Data Mining for Earth and Space Science Applications

R. Chen, C. Giannella, Krishnamoorthy Sivakumar, Hillol Kargupta

Collaborators: Kirk Borne, Ranga Myneni, Ashok Srivastava

Speaker: Hillol Kargupta

University of Maryland Baltimore County and AGNIK, LLC

http://www.cs.umbc.edu/~hillol

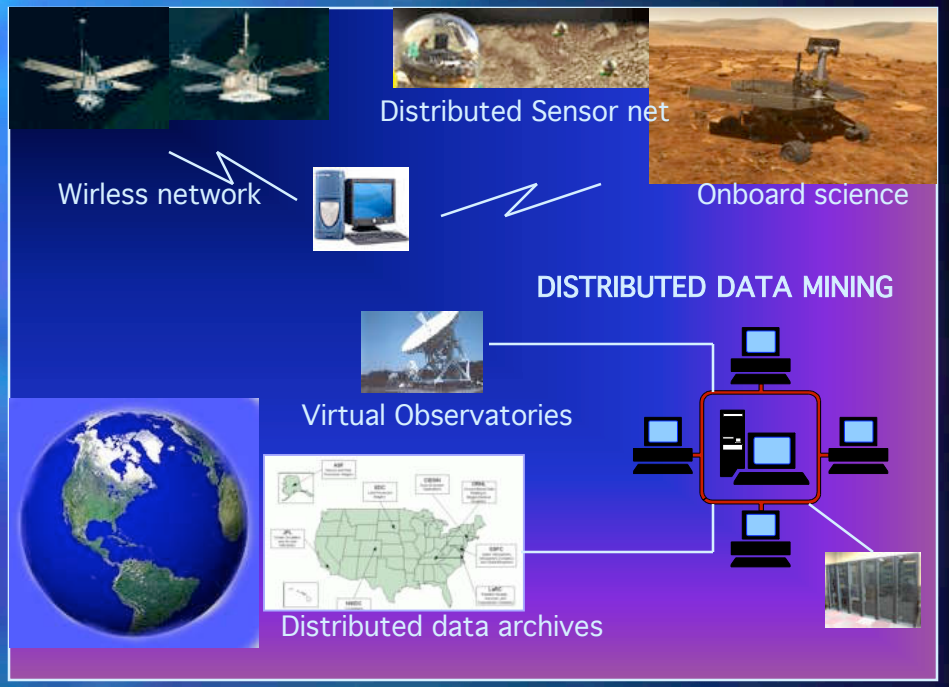
http://www.agnik.com, hillol@cs.umbc.edu

Roadmap

- Introduction
- Distributed data mining:
  - Overview
  - Distributed Bayesian network learning
  - Distributed decision tree learning
- Applications
  - Mining NASA DAO and NOAA data
  - Mining distributed virtual observatories
- Conclusions and future work

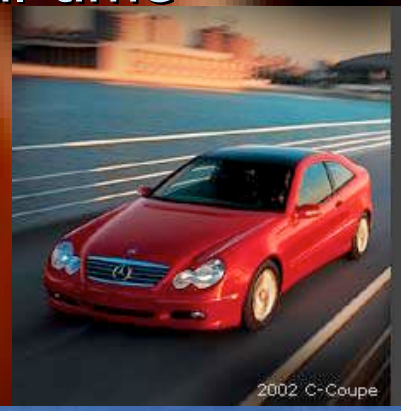
Overview

- Goal: Analyzing distributed heterogeneous data by properly utilizing distributed resources.
- Contributions:
  - Distributed computation of
    - Statistical aggregates
    - Decision trees
    - Bayesian networks.
  - Algorithms for monitoring distributed data streams.
  - Mining NASA/NOAA AVHRR data and the virtual observatory data.



Broader Impacts

- Mining Databases from distributed sites
  - Counter-terrorism, bioinformatics
- Monitoring Multiple time critical data streams
  - Monitoring vehicle data streams in real-time
  - Onboard science
- Analyzing Lightweight sensor webs
  - Limited network bandwidth
  - Limited power supply
- Preserving privacy
  - Security/Safety related applications



Why Bother?

x1	x2
4	1
4	5
6	8
1	4
7	1

x1	x3
4	5
1	6
5	9
2	10
7	4

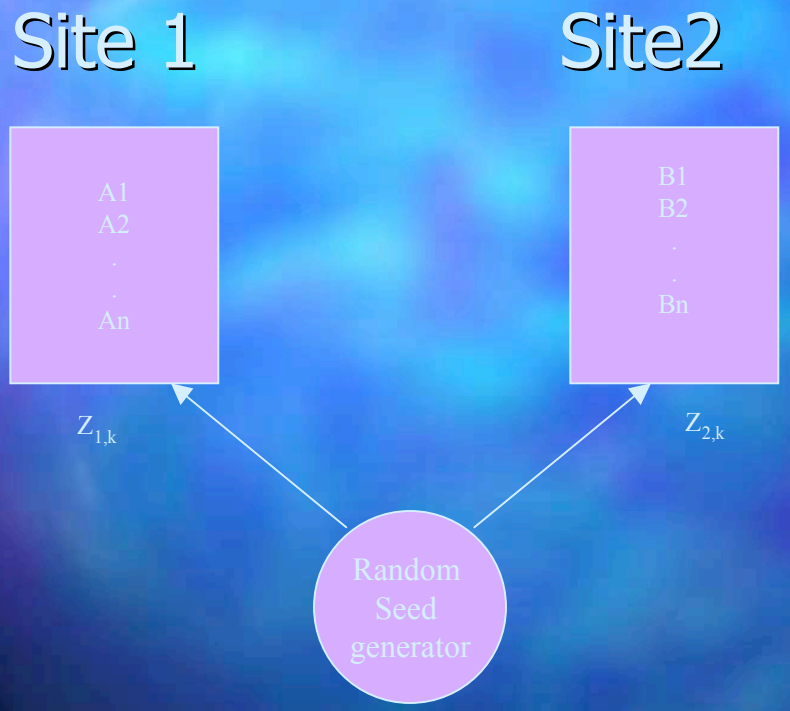
x1	x2	x3
4	1	5
4	5	5
1	4	6
7	1	4
…	…	…

- (Left) Data table at site 1.
- (Middle) Data table at site 2.
- (Right) Joined data table (based on the shared feature x_1) needed for centralized data mining systems.
- Problems:
  - Construction of the join is computationally expensive
  - Supporting repeated queries (e.g. for streams) may be too expensive for the communication-bandwidth.

Mining a Network of Virtual Observatories

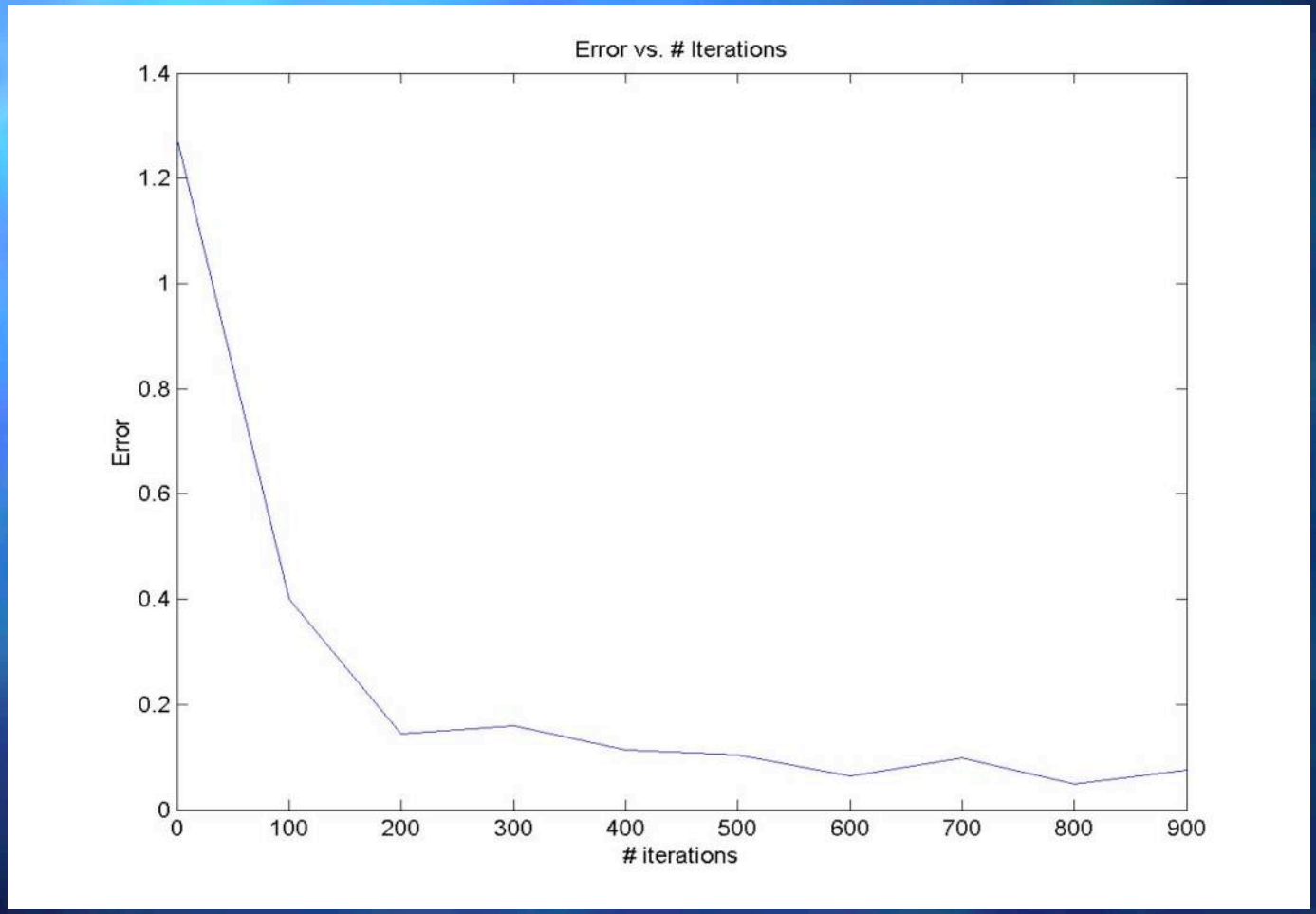
- The Sloan Digital Sky Survey (SDSS) and the 2MASS All-Sky Survey.
- Five filters from SDSS and three filters from 2MASS.
- Analyze data from SDSS and 2MASS:
  - Cluster the set of objects using attributes from both the observatories
  - Identify outliers
  - Learn classifiers

Distributed Randomized Inner Product Computation

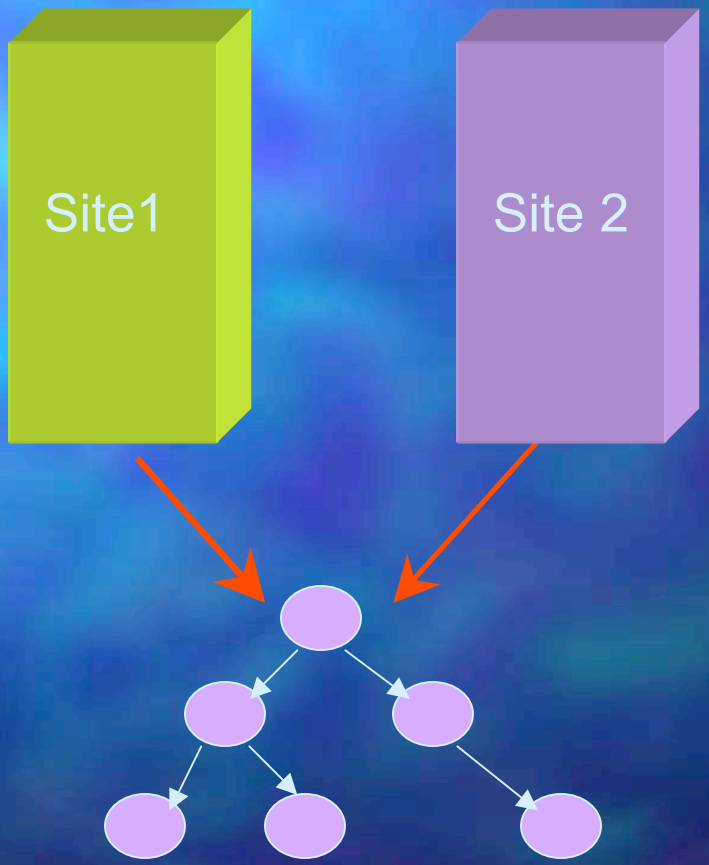


- Site 1 computes Z_{1k}
  - $Z_{1k}=A1.J_1+..+An.J_n$
  - $J_i \in \{+1,-1\}$ with uniform probability
- Site 2 calculates Z_{2k}
  - $Z_{2k}=B1.J_1+..+Bn.J_n$
- Compute $z_{1k}.z_{2k}$ for a few times and take the average

Relative Error vs. Communication Cost



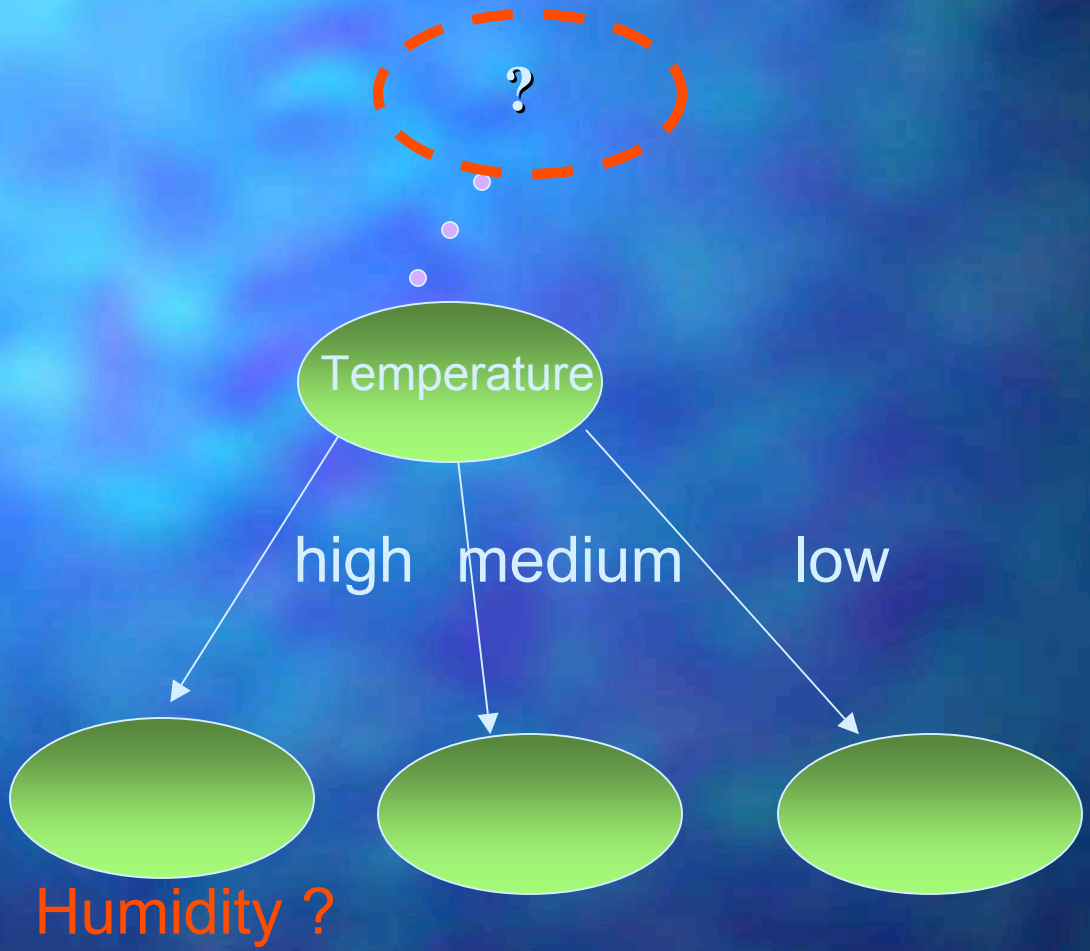
Decision Tree Induction From Vertically Partitioned Distributed Data



Heterogeneous DDM and Decision Trees

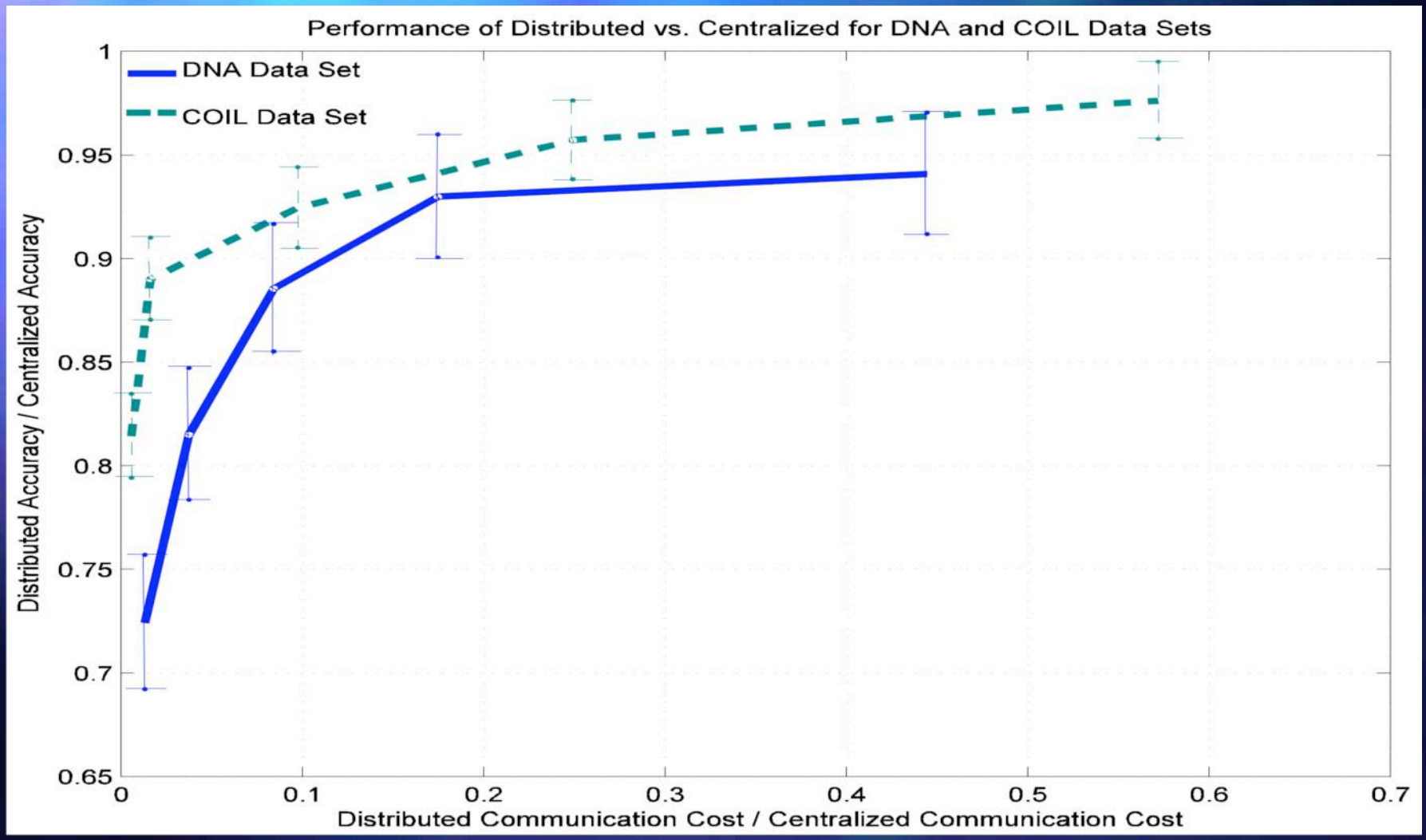
- Distributed Randomized Inner Product (DRIP) computation
- Computing information gain using DRIP.
- Information gain computation can be posed as an inner product computation problem.

Information Gain Computation and DRIP



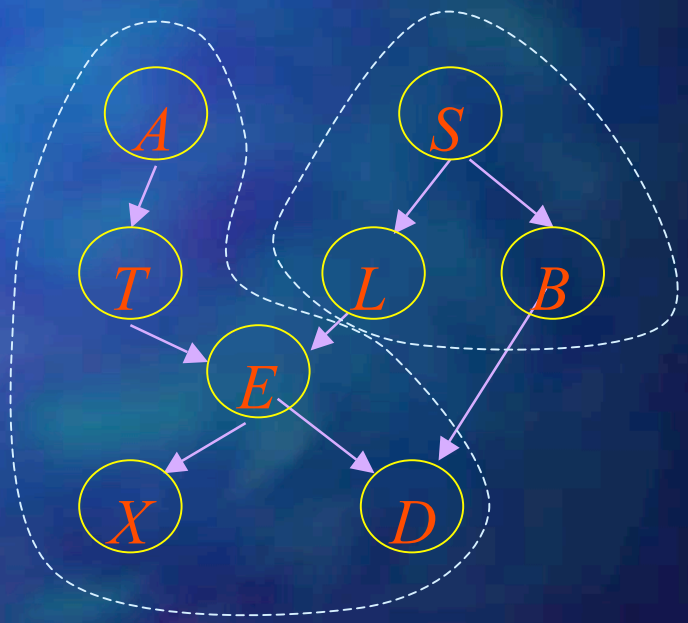
Information_Gain(Temp_high,Humidity) = ?

Experimental Results

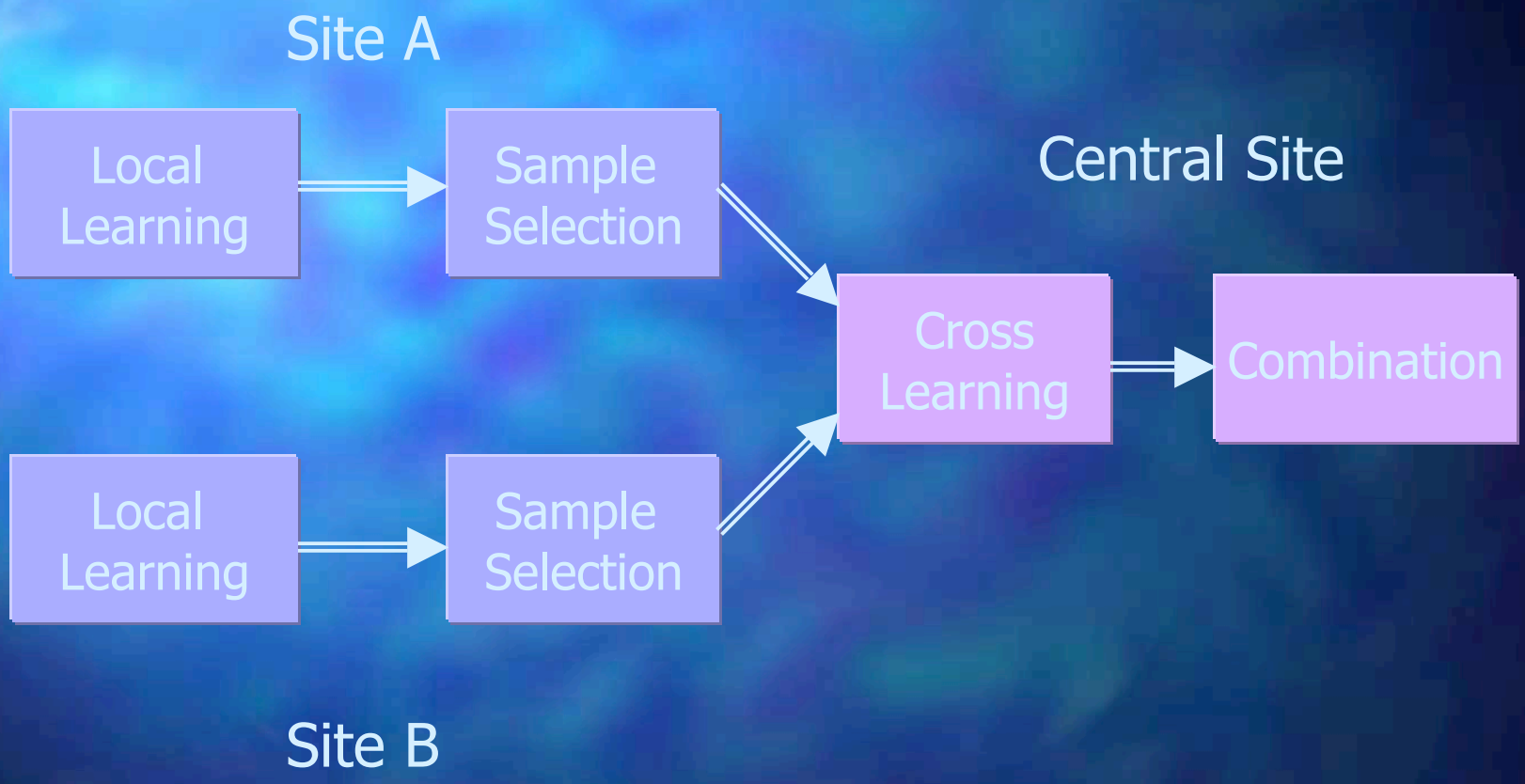


Distributed BN Learning

- A Bayesian network (BN) is a probabilistic graph model.
- Two problems: Structure and Parameter learning.

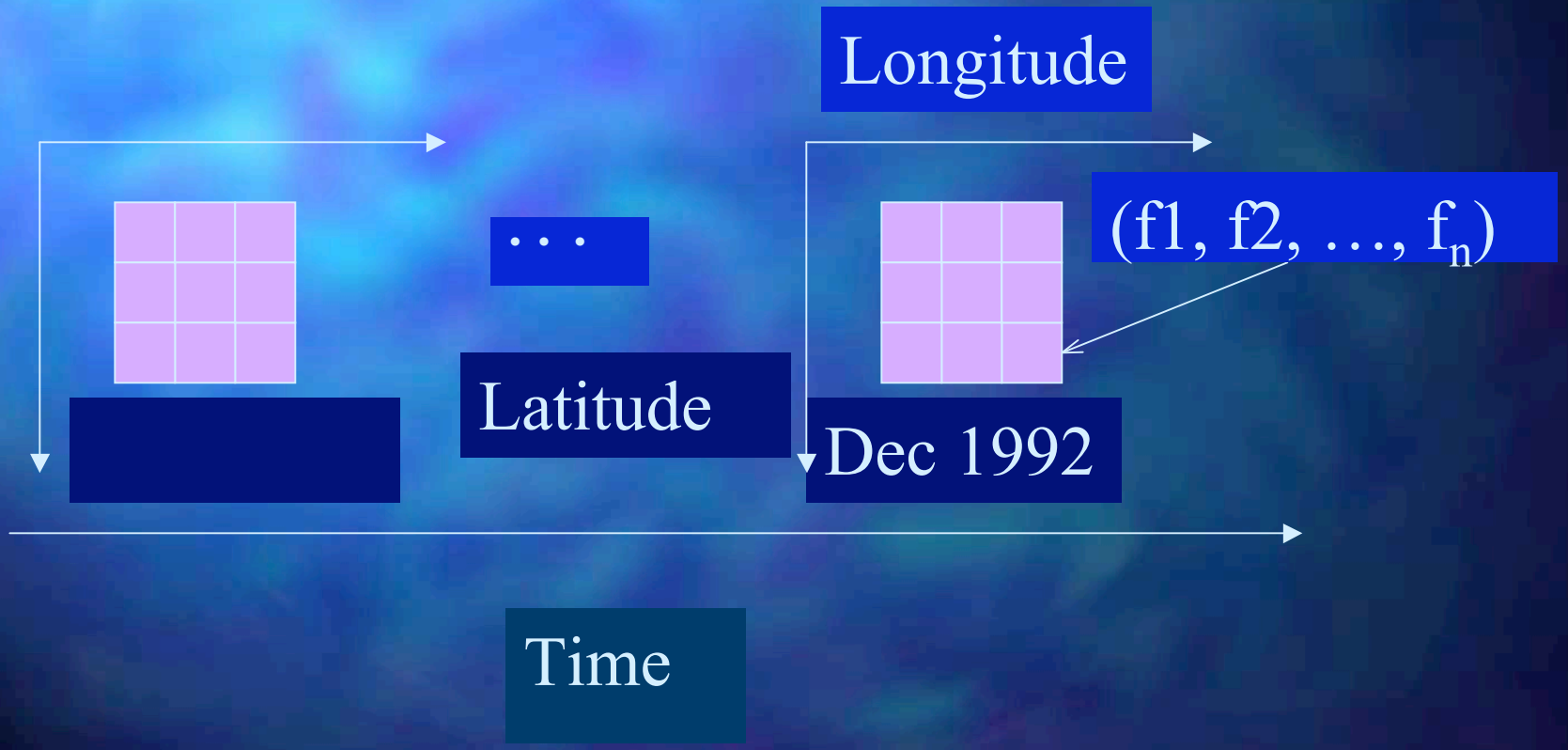


Collective BN Learning



NASA DAO/NOAA AVHRR Pathfinder Data Model

- Multi-dimensional time series data



Preprocessing

- Feature Selection
- Data Coordination
- Clustering: Segment grid points into local homogenous regions.
- Z score normalization
- Quantization

Feature Selection

- We used as many features as possible.
- Features with following characteristics were dropped.
  - Many missing values
  - Multi-layer features
  - Almost deterministic features
- Used 15 DAO and 7 NOAA features

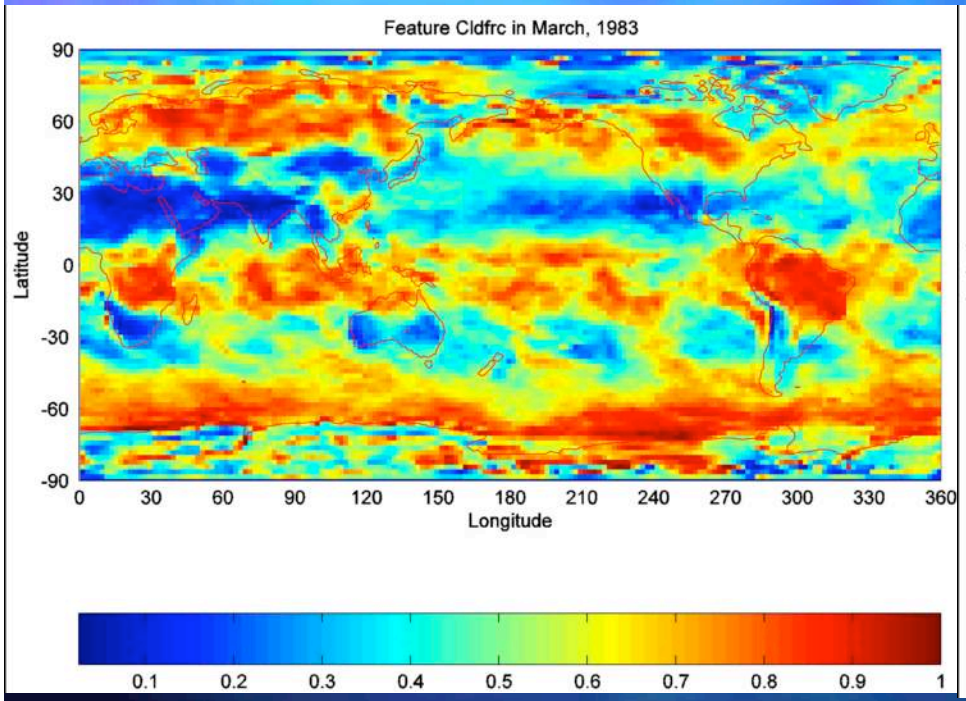
NASA DAO features

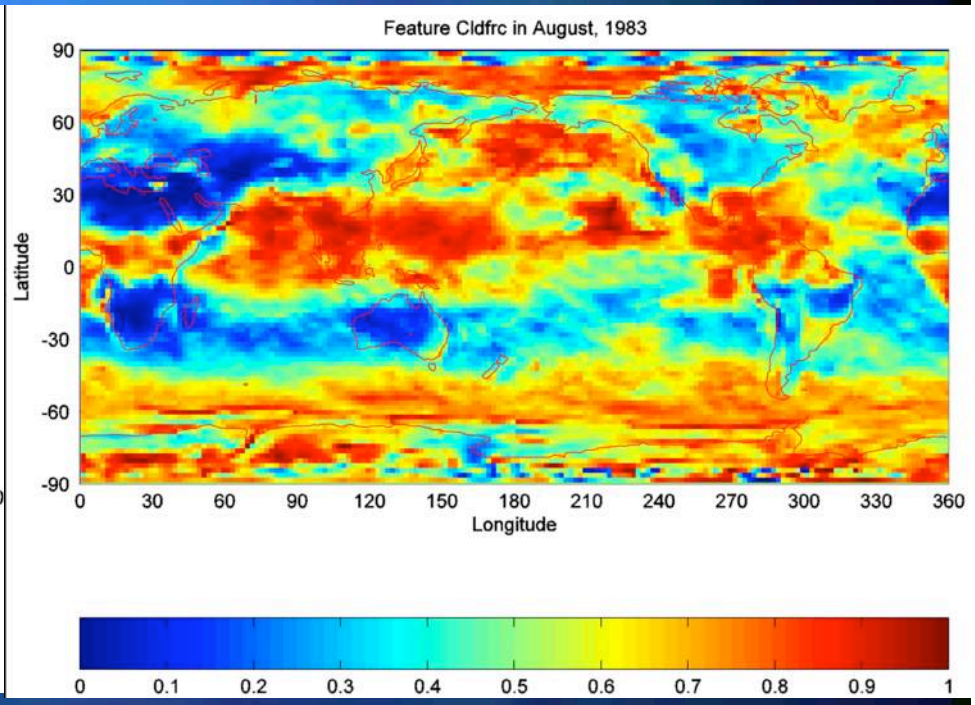
Index	Feature	Description
1	Cldfrc	2-dimensional total cloud fraction
2	Evaps	Surface evaporation
3	Olr	outgoing longwave radiation
4	Osr	outgoing shortwave radiation
5	Pbl	planetary boundary layer depth
6	preacc	total precipitation
7	qint	precipitable water
8	radlwg	net upward longwave radiation at ground
9	radswg	net downward shortwave radiation at ground
10	t2m	temperature at 2 meters
11	tg	Ground temperature
12	ustar	Surface stress velocity
13	vintuq	vertically averaged uwnd*sphu
14	vintvq	vertically averaged vwnd*sphu
15	winds	Surface wind speed

NOAA features

Index	Feature	Description
16	asfts	Absorbed Solar Flux total/day
17	olrcs day	Outgoing Long Wave Radiation clear/day
18	olrcs night	Outgoing Long Wave Radiation clear/night
19	olrts day	Outgoing Long Wave Radiation total/day
20	olrts night	Outgoing Long Wave Radiation total/night
21	tcf day	Total Fractional Cloud Coverage day
22	tcf night	Total Fractional Cloud Coverage night

Feature Cldfrc in March (left) and August (right), 1983





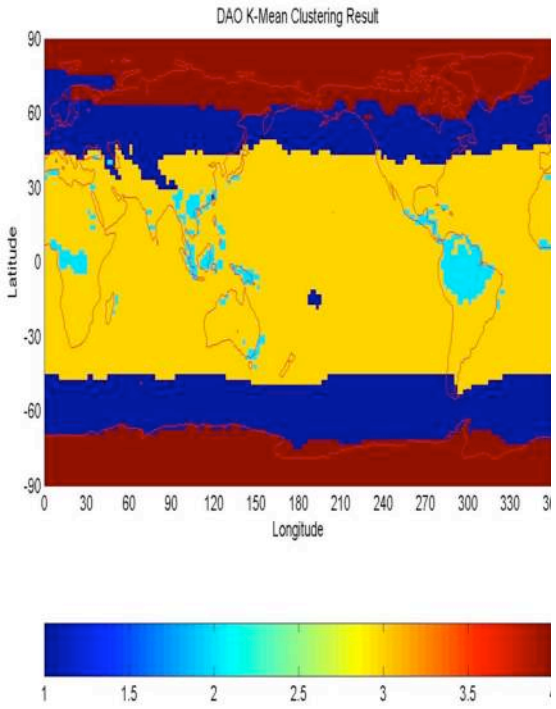
Coordination and Clustering

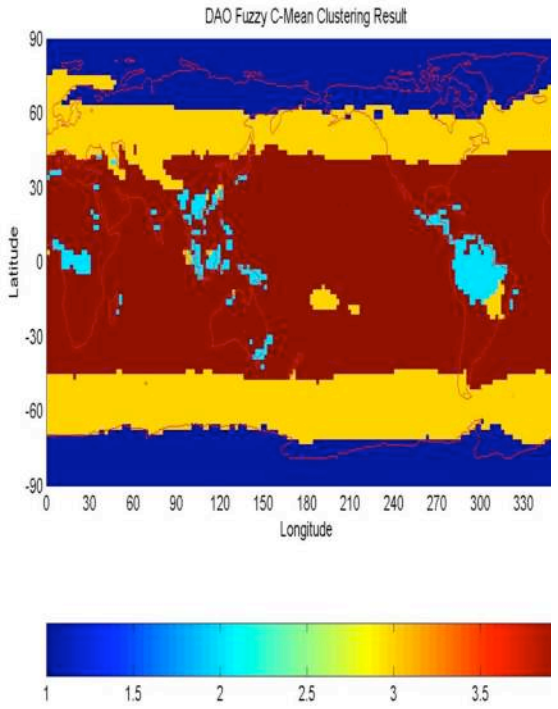
- **Coordination**: re-grid the NOAA dataset into DAO format.
- **Spatio-temporal Clustering**: Segment datasets into local homogenous regions in spatial and temporal domain.
- Each cluster is modeled using a Bayesian network.

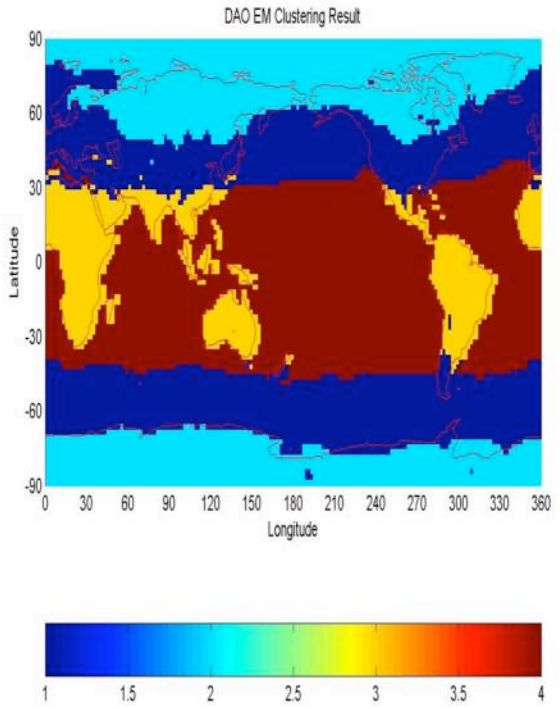
Spatio-temporal Clustering

- **Temporal clustering**: choose same month data.
- **Spatial clustering**
  - Average the data from same month. Get one frame of data in spatial domain.
  - Clustering: k-mean, fuzzy c-mean, and EM.

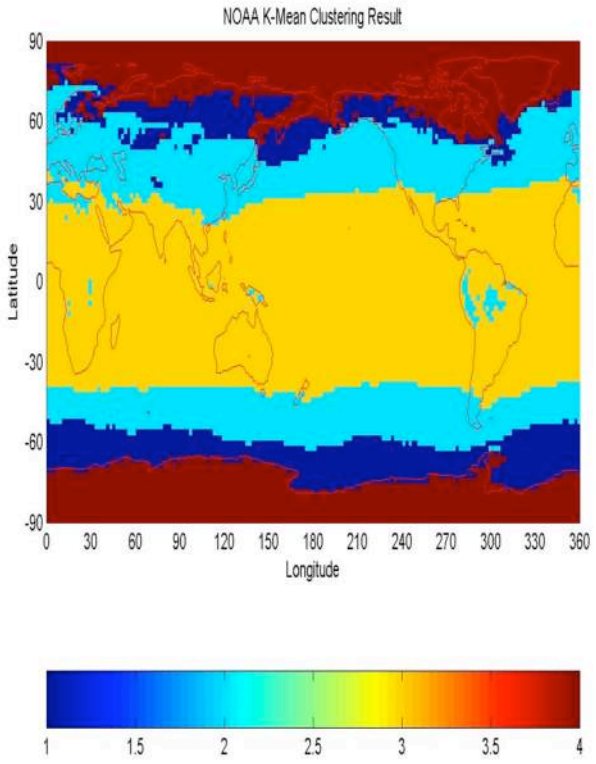
Clustering Results: DAO

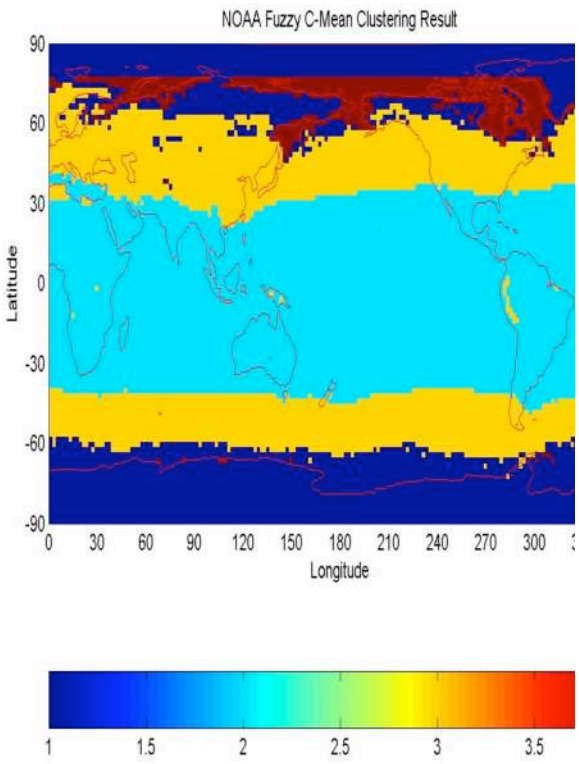


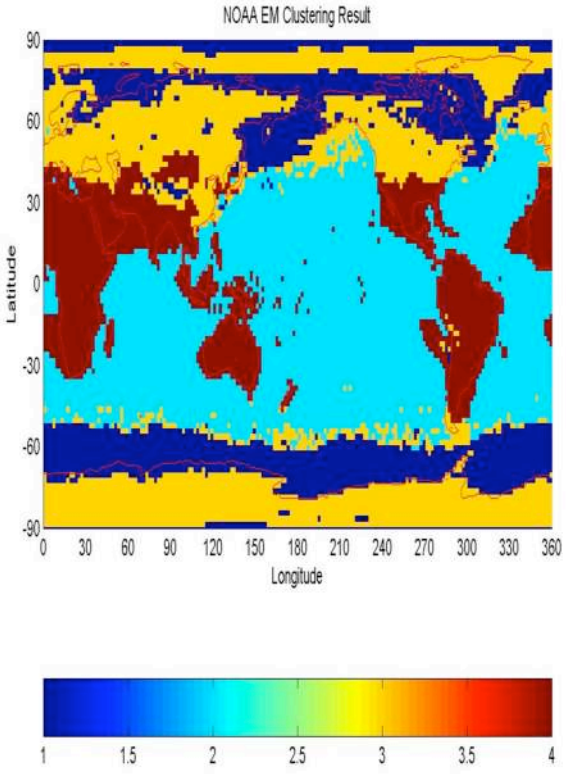




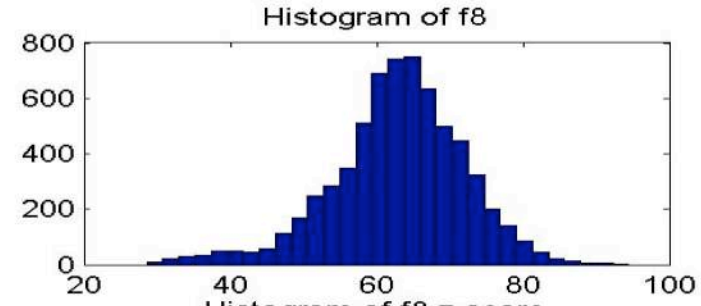
Clustering Results: NOAA

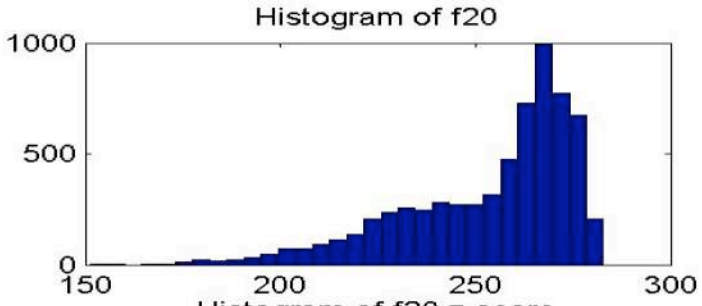


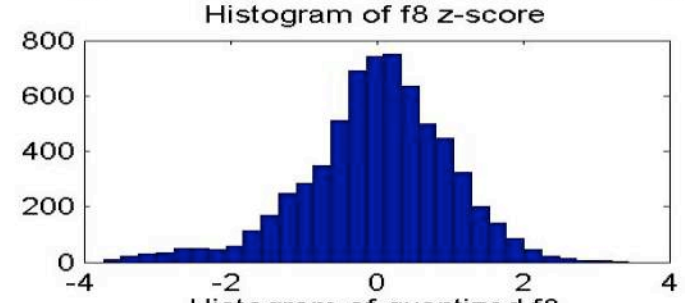


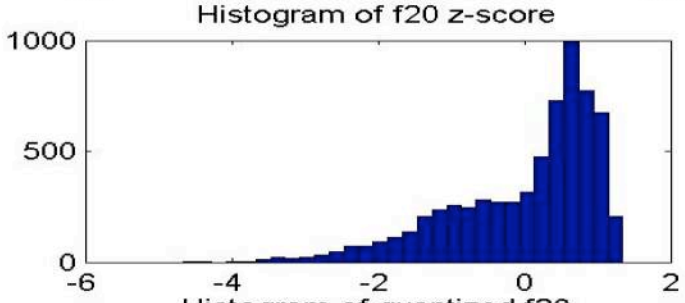


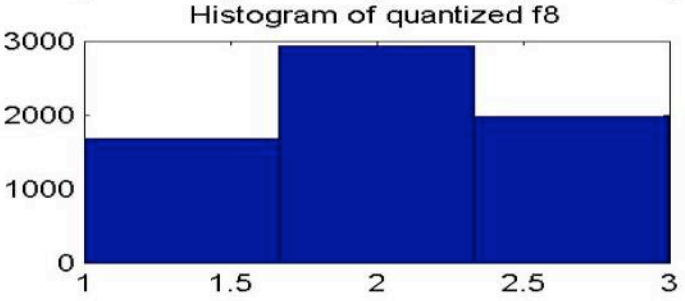
Quantization Results

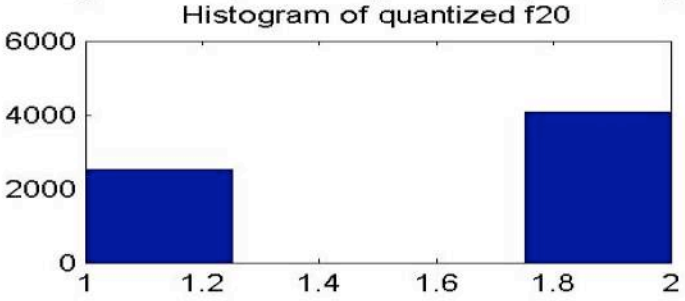










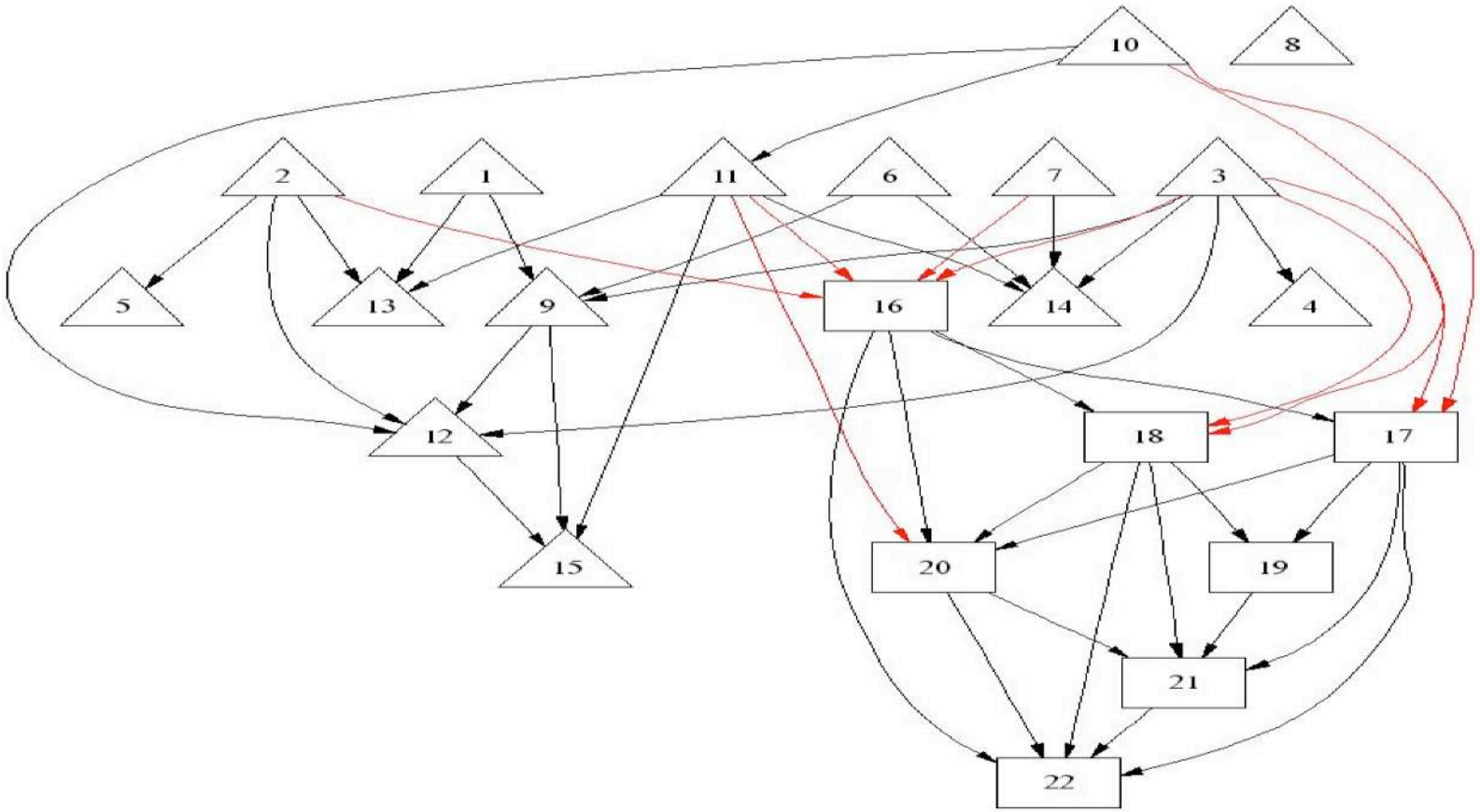


Bayesian network Learning Results

- Compare the Bayesian Networks:
  - B_{cntr} learnt using centralized method.
  - B_{coll} learnt using collective method.
- Metric: structure error = Number of missing links + Number of extra links.

Result

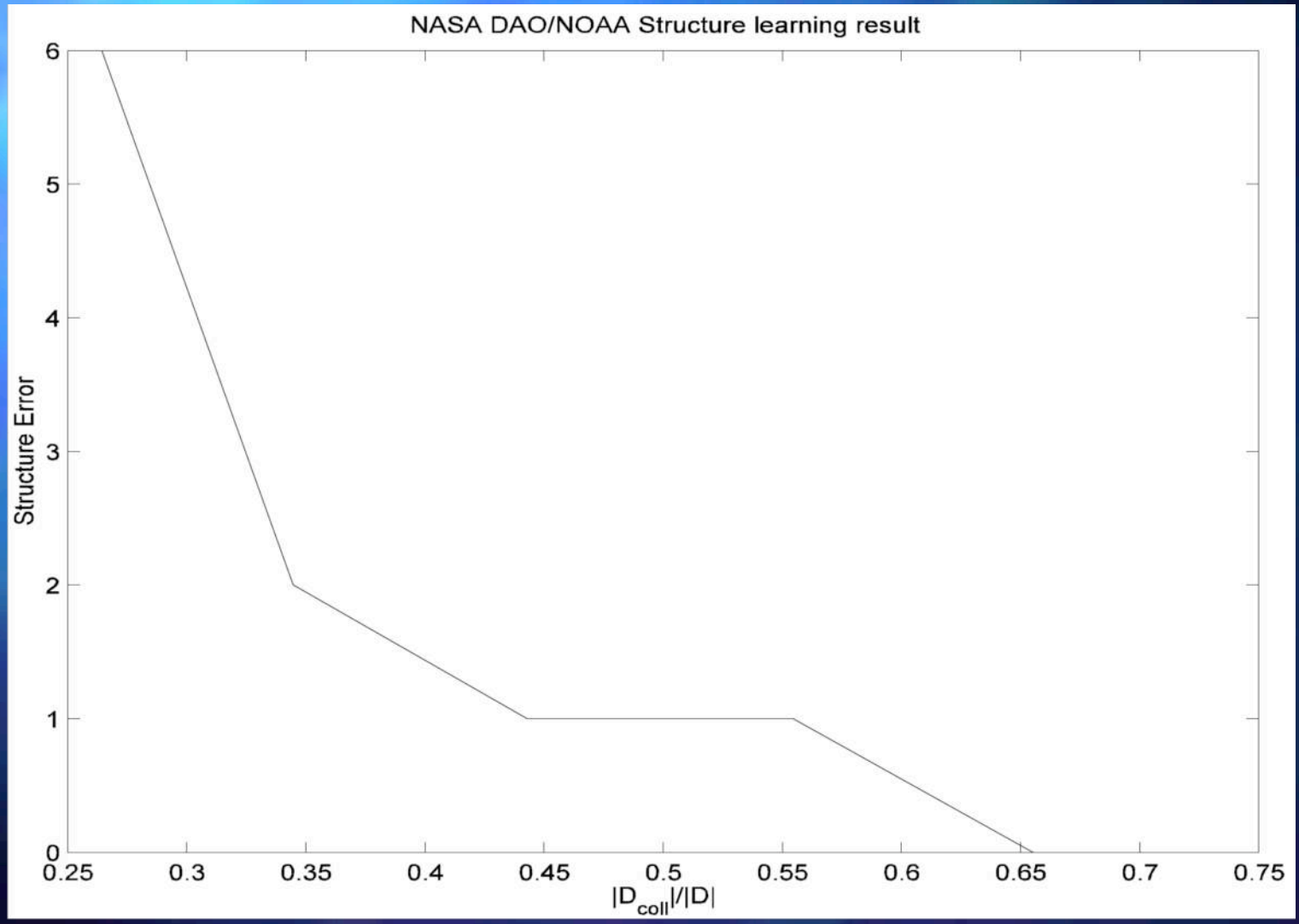
- B_{cntr} – 64 local links and 9 cross links.



Some of the Cross Links Between NOAA and DAO Attributes

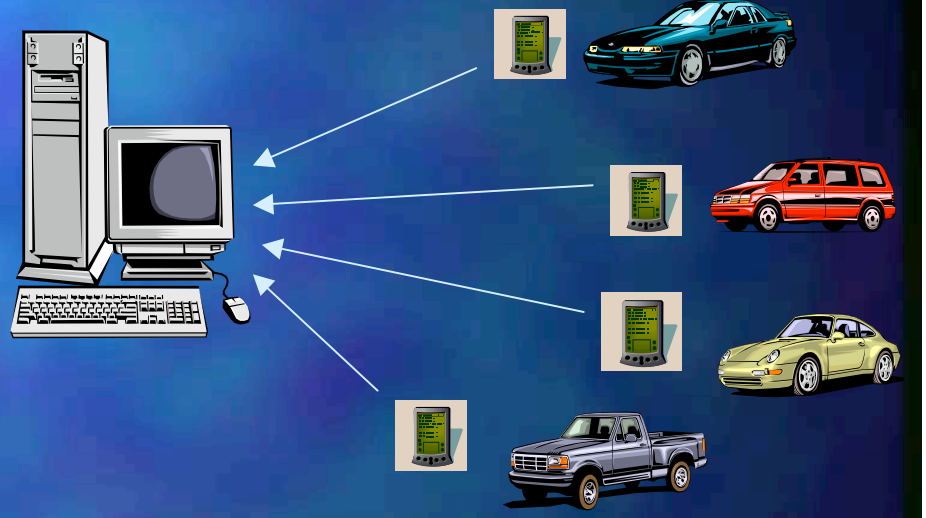
- Surface evaporation, Absorbed Solar Flux total/day
- outgoing longwave radiation, Absorbed Solar Flux total/day
- outgoing longwave radiation, Outgoing Long Wave Radiation clear/day
- outgoing longwave radiation, Outgoing Long Wave Radiation clear/night
- precipitable water, Absorbed Solar Flux
- temperature at 2 meters, Outgoing Long Wave Radiation clear/day
- temperature at 2 meters, Outgoing Long Wave Radiation clear/night
- Ground temperature, Absorbed Solar Flux total/day
- Ground temperature, Outgoing Long Wave Radiation total/night

NASA DAO/NOAA Structure Learning Results



Applications in Mobile Sensor Networks: Vehicle Data Stream Mining

- On-board Module:
  - Continuous data streams from the vehicle data bus
  - Onboard data stream mining
  - Communicates with a remote control station
  - Privacy management
- Central control station:
  - Data Management
  - Data mining
  - Communicates with the on-board modules over wireless networks
  - Privacy management



Conclusions

- Distributed data mining: A new way to do data mining in distributed environments
- Applications in mining,
  - Large distributed collection of repositories
  - Bandwidth/power constrained sensor networks
  - Privacy-sensitive multi-party data
  - Scalable time-critical analysis of data streams from different sources

Web site

- http://www.cs.umbc.edu/~hillol/nasap.html

Selected Publications

- Kargupta, H. and Park, B. (2004). A Fourier Spectrum-Based Approach to Represent Decision Trees for Mining Data Streams in Mobile Environments. IEEE Transaction on Knowledgeand Data Engineering, Volume 16, Number 2, pages 216--229.
- Chen, R., Sivakumar, K., and Kargupta, H. Collective Mining of Bayesian Networks from Distributed Heterogeneous Data. (accepted) Knowledge and Information Systems, 2002.
- Chen, R. and Sivakumar, K. A New Algorithm for Learning Parameters of a Bayesian Network from Distributed Data. (To appear) Proceedings of the IEEE International Conference on Data Mining, 2002, IEEE Press.
- Chen, R., Sivakumar, K., and Kargupta, H. Distributed Web Mining using Bayesian Networks from Multiple Data Streams. Proceedings of the IEEE International Conference on Data Mining, 75--82. IEEE Press.
- Chen, R., Sivakumar, K., and Kargupta, H. An Approach to Online Bayesian Learning from Multiple Data Streams, Proceedings of the Workshop on Mobile and Distributed Data Mining, PKDD2001.
- Kargupta, H. and Park, B. Mining Decision Trees from Data Streams in a Mobile Environment. Proceedings of the IEEE International Conference on Data Mining, 281--288. IEEE Press.
- Park, B. and Kargupta, H. Constructing Simpler Decision Trees from Ensemble Models Using Fourier Analysis, Proceedings of ACM SigMod DMKD'02 Workshops, Madison, WI (To appear).
- Ayyagari, R. and Kargupta, H. A Resampling Technique for Learning the Fourier Spectrum of Skewed Data, Proceedings of ACM SigMod DMKD'02 Workshops, Madison, WI (To appear).

Preprocessing

- **Clustering**: Chose a cluster that roughly corresponds to the rectangular region from (170W, 60S) to (90W, 0)
- **Z score normalization**

$$d_z = \frac{d - \mu}{\sigma}$$

- **Quantization**: Discretize the continuous feature value into discrete levels based on its histogram.
- After above steps, we get 12 datasets, one for each month (aggregated over years 1983-1992).

Quantization Results